論文 / 著書情報
Article / Book Information

論題(和文)	尤度比最大基準によるストリーム重み最適化を用いたマルチモーダル音声認識の性能評価
Title(English)	
著者(和文)	田村 哲嗣, 岩野 公司, 古井 貞熙
Authors(English)	Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会 2004年春季講演論文集, Vol. , No. 3-8-1, pp. 123-124
Citation(English)	, Vol. , No. 3-8-1, pp. 123-124
発行日 / Pub. date	2004, 3

尤度比最大基準によるストリーム重み最適化を用いたマルチモーダル音声認識の性能評価 *

◎田村 哲嗣　岩野 公司　古井 貞熙 （東工大）

1. はじめに

雑音環境下で頑健に音声認識を行う手法の一つとして，唇動画像の情報を利用するマルチモーダル音声認識が注目されている．我々はこれまでに，マルチモーダル音声認識における音響情報と画像情報の重み付け係数の最適化手法を提案し，実環境データを用いた認識実験で認識率の改善に成功している [1]．本論文では，この重み最適化手法と雑音適応手法（MLLR）とを組み合わせた実験や，MCE-GPD による最適化手法と比較した結果について報告する．

2. 尤度比最大基準による重み最適化

本研究では，音声認識時においてマルチストリーム HMM を使用している．いま，デコーダが出力した単語列を $w_1, w_2, \cdots, w_M$，単語 w_i $(1 \leq i \leq M)$ は時刻 $T_{i-1} \leq t < T_i$ で生起されたものとし，この時間の音響–画像観測系列を $\mathbf{O}^i$ とおく．このとき，単語 w のモデルに対する音響–画像平均対数尤度 $\bar{b}_w(\mathbf{O}^i)$ は，

$$\bar{b}_w(\mathbf{O}^i) = \lambda_{Aw}\bar{b}_{Aw}({\mathbf{O}_A}^i) + \lambda_{Vw}\bar{b}_{Vw}({\mathbf{O}_V}^i) \qquad (1)$$

と表される．ただし $\bar{b}_{Aw}({\mathbf{O}_A}^i)$，$\bar{b}_{Vw}({\mathbf{O}_V}^i)$ はそれぞれ音響観測系列 ${\mathbf{O}_A}^i$，画像観測系列 ${\mathbf{O}_V}^i$ に対する単語 w のモデルにおける音響，画像平均対数尤度，λ_{Aw}，λ_{Vw} は単語 w を構成する HMM における音響，画像ストリーム重みで，以下の制約がある．

$$\lambda_{Aw} + \lambda_{Vw} = 1 \ , \ 0 \leq \lambda_{Aw}, \lambda_{Vw} \leq 1 \qquad (2)$$

このストリーム重みに関して，我々は尤度比最大基準により決定する手法を提案している [1]．この手法では，適応データについて，正解（仮説）単語とその他の単語の対数尤度の差 $L_P(\Lambda)$ が最大となるよう，次式によりストリーム重み $\Lambda = \{\lambda_{Aw}\}$ を推定する．

$$L_P(\Lambda) = \sum_{i=1}^{M} \sum_{w \in W} \left\{ \bar{b}_{w_i}(\mathbf{O}^i) - \bar{b}_w(\mathbf{O}^i) \right\}^2 \rightarrow \max \quad (3)$$

ここで W は認識に用いる辞書 $(|W| = N)$ である．式 (3) より，任意の単語 $v \in W$ に対する λ_{Av} の変化分 $\Delta\lambda_{Av}$ は次のように求められる．

$$\begin{aligned} \Delta\lambda_{Av} &= P/Q \qquad (4) \\ P &= \sum_{i=1}^{M} \Big[\delta_{w_i=v} \cdot \Big\{ N\bar{b}_v(\mathbf{O}^i) - \sum_{w \in W} \bar{b}_w(\mathbf{O}^i) \Big\} \\ &\qquad + \delta_{w_i \neq v} \cdot \Big\{ \bar{b}_v(\mathbf{O}^i) - \bar{b}_{w_i}(\mathbf{O}^i) \Big\} \Big] \\ Q &= \sum_{i=1}^{M} \Big[\delta_{w_i=v} \cdot N\bar{f}_v(\mathbf{O}^i) + \delta_{w_i \neq v} \cdot \bar{f}_v(\mathbf{O}^i) \Big] \end{aligned}$$

表 1: 実験条件（使用特徴量・適応の有無）

	使用特徴量	MLLR	重み最適化
(A)	音響のみ	×	×
(B)	音響–画像	×	×
(C)	音響–画像	×	○
(D)	音響–画像	○	×
(E)	音響–画像	○	○

$$\bar{f}_w(\mathbf{O}^i) \ = \ \bar{b}_{Aw}({\mathbf{O}_A}^i) - \bar{b}_{Vw}({\mathbf{O}_V}^i)$$

ここで δ_x は x が真のとき 1，偽のとき 0 を返す関数である．式 (4) により全ての $v \in W$ に対する $\Delta\lambda_{Av}$ を求め，その後 λ_{Av} を更新する．このサイクルを繰り返すことで最適なストリーム重みを推定できる．

3. MLLR と組み合わせた認識実験

尤度比最大基準によるストリーム重み最適化と，雑音適応に広く用いられている MLLR 適応を組み合わせ，マルチモーダル音声認識システム [1] により認識実験を行った．

3.1. データベース

学習データにはクリーン環境で収録した男性話者 11 名の，テストデータには高速道路走行中の車内で収録した，6 名の数字連続読み上げデータを用いた [2]．各話者は 2〜6 桁の数字を，学習データでは 250 個，テストデータでは 115 個発声している．

3.2. 実験条件

実験条件を表 1 に示す．(A) はベースライン，(B) ではストリーム重み最適化を行わず全モデルに同じストリーム重みを用いた．(C)〜(E) では，MLLR と重み最適化の，片方あるいは両方を適用した．適応・最適化は教師なしで行い，使用するラベルは (B) の認識結果を用いた．MLLR では話者ごとにバッチ適応により，音響ストリームの正規分布の平均・分散を適応した．重み最適化では，音響ストリーム重みの初期値を 1.0 とし，全テストデータを用いて推定を行い，繰り返し演算回数は 50 回とした．(E) では，MLLR 適応の後に重み最適化を行った．

3.3. 実験結果

図 1 に (A)〜(E) の実験条件に対する数字正解精度を示す．横軸は (B) で用いた音響ストリーム重み，縦軸は数字正解精度である．またそれぞれの条件での最も高い数字正解精度を表 2 に示す．(A) と比べると，ストリーム重み最適化を行った (C) では正解精度では約 14%改善，誤り率は約 36%削減でき，重み最適化に

---

* Evaluation of multi-modal speech recognition using a stream-weight optimization method based on a likelihood-ratio maximization criterion,
by Satoshi Tamura, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology).

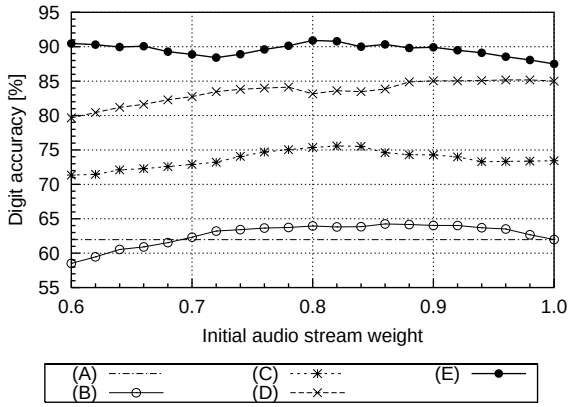

図 1: 各種実験条件における認識結果

表 2: 各種実験条件における最高数字正解精度

(A)	(B)	(C)	(D)	(E)
62.0%	64.2%	75.6%	85.2%	91.0%

よって認識性能が大幅に向上した．また (D) と (E) を比べると，約 39%誤り率が削減されており，MLLR 適応の有無に関わらずほぼ同じ誤り率の削減を達成できた．最終的に (A) と (E) から，MLLR と重み最適化を組み合わせることで，約 29%の数字正解精度の改善，約 76%の誤り率の削減を達成した．

4. MCE-GPD 法との比較実験

次に，尤度比最大基準によるストリーム重み最適化と MCE-GPD による最適化の比較を行った．

4.1. MCE-GPD 法

ストリーム重みを推定する手法としては，最小分類誤り基準（MCE）による方法がよく用いられている [3, 4]．この手法では，誤分類測度 $d_w(\mathbf{O}^i;\Lambda)$ に対して損失関数 $l_w(\mathbf{O}^i;\Lambda)$ を定め，式 (7) のように $L_M(\Lambda)$ を最小化することで，ストリーム重み Λ を推定する．

$$d_w(\mathbf{O}^i;\Lambda) = -\bar{b}_w(\mathbf{O}^i) + \frac{1}{\eta}\log\left(\sum_{w\in W} e^{\eta\bar{b}_w(\mathbf{O}^i)}\right) \quad (5)$$

$$l_w(\mathbf{O}^i;\Lambda) = 1 \Big/ \left\{1 + e^{-\alpha d_w(\mathbf{O}^i;\Lambda)}\right\} \quad (6)$$

$$L_M(\Lambda) = \sum_{i=1}^{M} l_{w_i}(\mathbf{O}^i;\Lambda) \to \min \quad (7)$$

ただし $\alpha, \eta > 0$ である．式 (7) から，Λ は GPD 法を用いて次式のように繰り返し演算で求められる．

$$\Lambda_{k+1} = \Lambda_k - \epsilon_k E \nabla L_M(\Lambda_k) \quad (8)$$

ここで k は繰り返し回数，E は単位行列，ϵ_k は単調減少する正の数列である．

4.2. 実験条件

テストデータの時間情報つき正解ラベルを用いた教師あり条件で，(F) 提案手法と (G) MCE-GPD 法によりストリーム重み最適化を行った．いずれも重み初期値は 1.0 とし，推定時の繰り返し回数は 50 回とした．(G) の制御パラメータについては，$\alpha = 1.4, 1.6, \cdots, 2.2$, $\eta = 0.8, 1.0, 1.2, 1.4$，$\epsilon_k = 1/k$ とし，α，η は重み最適化後に認識率が最も高かった値を事後的に選択した．

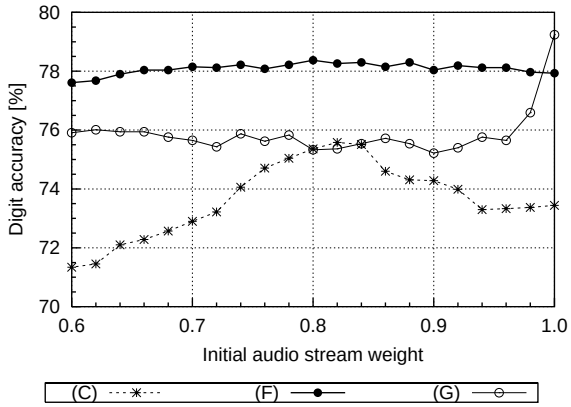

図 2: 尤度比最大基準と MCE-GPD による最適化で得られたストリーム重みによる認識結果

4.3. 実験結果

図 2 に，(F)，(G) に対する数字正解精度を，(C) の結果とあわせて示す．横軸は時間情報ラベル生成に用いた音響ストリーム重み λ_A，縦軸は数字正解精度である．図より，$\lambda_A = 1.0$ では (G) の方が高いが，それ以外では (F) の方が高い性能を示した．MCE-GPD 法には適切な制御パラメータを設定することが難しいという問題もあり，この点からも，我々の手法は実用面において有利である．また (C) と (F) の比較より，提案手法における教師ありと教師なしの差は約 3%と小さいことから，教師なしでの尤度比最大基準による最適化手法の有効性を確認できた．

5. まとめ

本論文では，我々が提案している尤度比最大基準によるストリーム重み最適化手法について，さまざまな条件で性能評価を行った．MLLR による雑音適応と組み合わせることにより，教師なしで約 29%と正解精度が大幅に改善した．また教師あり条件で MCE-GPD による最適化手法との比較を行ったところ，全体的に MCE-GPD よりも高い性能を示し，提案手法の実用性，頑健性を確認した．今後の課題としては，結果統合法の検討，重み最適化手法のオンライン音声認識など他のタスクへの適用などが挙げられる．

謝辞

本研究は NTT ドコモ株式会社の援助を受けて行われました．ここに深く感謝いたします．

参考文献

[1] 田村 哲嗣, 岩野 公司, 古井 貞熙, “マルチモーダル音声認識における音響・画像特徴の融合法に関する検討,” 2003 年秋季音講論, 3-6-11, pp.123-124 (2003-9).

[2] 田村 哲嗣, 岩野 公司, 古井 貞熙, “実環境におけるマルチモーダル音声認識の評価,” 2002 年春季音講論, 3-5-5, pp.151-152 (2002-3).

[3] 宮島 千代美, 徳田 恵一, 北村 正, “最小誤り学習に基づくバイモーダル音声認識,” 2000 年春季音講論, 1-Q-14, pp.159-160 (2000-3).

[4] K.Kumatani and S.Nakamura, “Audio-visual speech recognition based on optimized product HMMs and GMM based-MCE-GPD stream weight estimation,” IEICE Trans. on Information and Systems vol.E86-D, no.3, pp.454-463 (2003-3).